

CHAPTER 14: TAAS WRITING ASSESSMENT

The TAAS writing tests at Grades 4 (both the English and Spanish versions), 8, and exit level consist of multiple-choice test items and a written composition. While the multiple-choice items test sentence structure, usage, and language mechanics, the written composition is a direct measurement of the student's ability to synthesize the component skills of writing; that is, the composition task requires the student to express ideas effectively in writing. To do this, he or she must be able to respond to a specific purpose and audience, to organize ideas clearly, to maintain a consistent control of written language, and to generate and develop ideas in a way that allows the reader to understand what the writer is attempting to say.

A process called focused holistic scoring is used to assess TAAS written compositions. The scoring system is “holistic” in that the piece of writing is considered as a whole; it is “focused” in that the piece of writing is evaluated according to preestablished criteria: response to purpose/audience, organization, language control, and elaboration. These criteria correspond to the first four writing objectives and are used to determine the effectiveness of each written response. Each TAAS response is evaluated according to the extent to which it reflects mastery of these four objectives. The response is scored on a scale of 1 (low) to 4 (high). A rating of 0 is assigned to compositions that are nonscorable. In addition, all responses that receive a rating of 0 or a score of 1 are evaluated analytically to determine why they are unsuccessful. This information is provided to districts in two forms: analytic designation(s) on the *Confidential Student Report* for individual students and aggregations of analytic designations on the *Written Composition Analytic Information Summary Report* for individual campuses and districts.

RUBRIC DEVELOPMENT

Generic scoring rubrics for the type of writing assessed and the grade level tested have been used throughout the TAAS program to ensure that the same standards are maintained across years and across prompts. The scoring rubrics are based on the following criteria. The student should be able to

- respond appropriately to a specific purpose and audience;
- employ one or more consistent organizational strategies;
- exhibit control of written language; and
- effectively develop the composition's central idea(s).

WRITTEN COMPOSITION SCORING

The written composition component of the TAAS writing test is scored by readers who are organized into teams that are coordinated by team leaders. All readers and team leaders receive extensive training on materials related specifically to the writing prompts for each administration. Through various training and practice sessions, during which readers score papers that have predetermined scores, readers are required to demonstrate complete understanding of the scoring process and must agree with the a priori scores on the training papers. During the actual scoring of student responses, validation sets of papers are systematically distributed to readers to monitor whether they are consistently applying the criteria and whether there is any drift from the “true” score. Released scoring guides, which include rubrics and sample student responses, are sent to all Texas school districts and regional education service centers every August.

TRAINING MATERIALS AND MONITORING OF PROJECT

The spring 2002 scoring of the written composition portion of the TAAS writing test took place at three different NCS Pearson Performance Scoring Center locations. Grade 4 English responses were scored in Lawrence, Kansas; Grade 8 responses were scored in Dallas, Texas; and exit level and Grade 4 Spanish responses were scored in Austin, Texas. Preparation for TAAS focused holistic scoring proceeded according to the following plan. NCS Pearson senior project staff and TEA staff, after independently scoring a selection of field-test responses, met in range-finding sessions in December 2001, January 2002, and February 2002 to discuss those responses and to assign “true” scores—both holistic and analytic—to the responses.

After range finding, the responses and their “true” scores were provided to the scoring directors of the respective grades, who worked with senior staff in Austin, Dallas, and Lawrence to assemble the materials for training team leaders and readers. Following TEA approval of the scoring guides and “split sets,” which contain responses that are “close calls” between one score point and the next, the scoring directors assigned the remaining prescored responses to training sets and qualifying rounds used to certify team leaders and readers. After the scoring directors annotated the guide responses, all training materials were photocopied.

TEA staff monitored training in Austin, Dallas, and Lawrence, selected validity papers (see “Validity Packets” on p. 75) and worked with senior staff and the Austin analytic coordinators in preparation for analytic scoring. Working with the 1/2 score verification specialist/analytic coordinator, TEA staff selected some “live” papers to include in 1/2 line and analytic training sets for the Austin exit level projects. Throughout the scoring process, senior staff served as on-site monitors in Austin, Dallas, and Lawrence.

Three exit level retests are scheduled annually. Two of these retests—those in October and July—are open to any exit level student who has previously failed one or more subject-area TAAS tests. The late April/early May retest is restricted to out-of-school students and those students who have enough credits to graduate but who have failed one or more subject-area TAAS tests.

For exit level retest scoring sessions in Austin, TEA and senior NCS Pearson staff selected validity papers (see “Validity Packets”) and monitored training and scoring.

MANAGEMENT-LEVEL STAFF

The NCS Pearson contract with TEA stipulates that all management-level staff at the scoring centers be approved by TEA. Accordingly, NCS Pearson submitted lists of scoring director candidates to TEA for approval, and TEA approved all candidates. All staff had extensive experience with the TAAS program and with numerous other large-scale state writing assessments.

RECRUITMENT OF READERS AND TEAM LEADERS

Requirements for readers included a bachelor’s degree, preferably in English, education, or a related field; teaching experience was preferred. In addition, all applicants were required to write an essay and complete a proofreading exercise. Those applicants interested in scoring the Grade 4 Spanish written composition were required to complete a Spanish decoding/translation exercise.

In February 2002 team leaders were selected by scoring directors. Team leaders were chosen from experienced readers and/or past team leaders. This selection process was based on team leaders’ understanding of the criteria, their ability to apply the criteria consistently and accurately, their ability to articulate the criteria, and their demonstration of leadership skills. Except for the “floating” team leader, team

leaders were assigned teams of 10 to 12 readers. The floating team leader assisted the scoring director with various administrative and quality-control activities.

In Austin a total of 580 people were involved in the Grade 4 Spanish and exit level scoring projects. For Austin's fall 2001 and summer 2002 retest scoring, all employees were selected from those who had participated in previous TAAS scoring efforts. Some 70 employees contributed to the success of the October 2001 project. Nineteen staff members worked on the May 2002 project, and 56 worked on the July 2002 project.

In Lawrence a total of 232 people were involved in the spring 2002 scoring project for Grade 4 English. In Dallas a staff of 258, including readers, team leaders, scoring directors, and coordinators, worked to complete the spring 2002 project for Grade 8.

ADMINISTRATIVE ARRANGEMENTS

The NCS Pearson scoring centers each had a day shift and a night shift throughout the focused holistic scoring of the spring 2002 TAAS responses. The analytic scoring and the score verification of failing exit level compositions (in Austin only) were accomplished during the day shift.

Scoring of the May 2002 retest in Austin took place within a one-week time frame beginning May 7, 2002. The daily schedule for this retest, as well as for the October 2001 and July 2002 exit retests, was the same as in spring 2002.

TRAINING

TRAINING MATERIALS: HOLISTIC SCORING

Training materials consist of guides, split sets, training sets, and qualifying sets. Following are descriptions of each of these training materials.

Guides

There were a total of sixteen student responses: four annotated anchor responses representing each score point in order, from 1 to 4.

Split Sets

Each split set contained four "close call" papers that defined the "line" between two score points. There was one split set for the 1/2 line, one for the 2/3 line, and one for the 3/4 line.

Training Sets

Training Sets A, B, and C each contained ten randomly mixed responses representing score points 1–4. Training Set D contained 15 randomly mixed responses representing score points 1–4.

Qualifying Sets

Each of the three qualifying sets contained twenty randomly mixed responses representing score points 1–4.

These training materials included 133 responses for each grade level tested. Following is a breakout by score point:

Grade 4	Grade 4 (Spanish)	Grade 8	Exit Level (Spring)
29 — 1s	32 — 1s	29 — 1s	29 — 1s
38 — 2s	36 — 2s	38 — 2s	38 — 2s
38 — 3s	36 — 3s	38 — 3s	38 — 3s
+28 — 4s	+29 — 4s	+28 — 4s	+28 — 4s
<hr/> 133	<hr/> 133	<hr/> 133	<hr/> 133

The October 2001, May 2002, and July 2002 exit level retest scoring sessions used training sets with a similar composition, except that they contained slightly more 1s and slightly fewer 4s to reflect more accurately the types of responses the readers would be scoring in those sessions.

TRAINING SET CONTENTS: EXIT LEVEL VERIFICATION SCORING

These training materials consisted of a guide using the 1s and 2s from the holistic guide and split set, along with ten-paper sets that constituted the rest of the 1s and 2s from the holistic training and qualifying sets. Additional responses found in the “live” papers (approved by the scoring director[s], the coordinator[s] of the analytic readers and 1/2 score verification specialists, and TEA staff) rounded out the sets.

TRAINING SET CONTENTS: ANALYTIC SCORING

The following analytic categories were used to explain why responses that received a rating of 1 or 0 (nonscorable responses) were unsuccessful.

Holistic Score Point 1

Analytic Categories

- Lacks clarity
- Lacks language control
- Lacks organization/structure
- Lacks support/elaboration
- Drifts from specified purpose
- Uses wrong purpose
- Drifts from specified topic

Holistic Score Point 0

Analytic Categories

- Off topic
- Indecipherable response
- Insufficient response to specified task

Training materials consisted of an eight-paper guide and a four-paper split set (from the holistic guide and 1/2 split set) demonstrating the successful/unsuccessful holistic response line, a ten-paper “Analytics” guide, and an explanation of the analytic categories and the numerical system used to assign the appropriate category or categories to each response. In addition, readers received four ten-paper training sets representing the various analytic categories and allowable combinations of categories.

TEAM LEADER TRAINING

Effective reader training for the October 2001 and the spring and July 2002 TAAS scoring sessions relied to a great extent on having knowledgeable, flexible team leaders. Because of the large number of readers, the scoring directors depended on team leaders to conduct small-group discussions with their teams; consequently, team leader training was critical to the success of the scoring effort.

The only scoring sessions for which team leaders were not used was May 2002. Because this retest was restricted to seniors and out-of-school students, the number of responses scored was small. Therefore, the scoring director conducted all training and team leaders functioned as readers.

Throughout their training, team leaders (joined by the analytic coordinators and, for exit level, the score-verification specialist coordinator) were encouraged to ask questions and to discuss any problems they had with the guide and the training sets. They were required to annotate their sets of training papers and to practice explaining their annotations to the rest of the group. Through this procedure the team leaders developed confidence in their ability to explain why a paper had been given a particular score. The guidelines for team leader and reader training were essentially the same. The specific steps were as follows:

1. Present the prompt in the exact form in which it was administered.
2. Read and explain the introduction section of the scoring guide.
3. Present a “good” paper (one that received a 4) from the scoring guide.
4. Proceed through the guide in the following manner:
 - a. Read and explain the score point 1 rubric. Read and discuss each annotated score point 1 paper.
 - b. Read and explain the score point 2 rubric. Read and discuss each annotated score point 2 paper. Read and discuss the 1/2 Split Set.
 - c. Read and explain the score point 3 rubric. Read and discuss each annotated score point 3 paper. Read and discuss the 2/3 Split Set.
 - d. Read and explain the score point 4 rubric. Read and discuss each annotated score point 4 paper. Read and discuss the 3/4 Split Set.
5. Score and discuss Training Sets A, B, C, and D.

After completing all the training sets, the team leaders took the qualifying sets. Regardless of whether a team leader scored well enough on Set I to qualify, he or she still took Set II. Taking both sets was important, since team leaders were responsible for working directly with readers; consequently, it was necessary for them to understand both qualifying sets. Set III was reserved for additional training as appropriate.

READER TRAINING

Before training began, readers signed their contracts and nondisclosure forms, and TEA representatives made introductory remarks.

The scoring director discussed the prompt, introduced the guide, and then explained each score point to the entire group of readers. Subsequently, readers broke into teams to score and discuss the papers in the training sets. The scoring director “floated” from team to team, listening to the team leaders’ explanations and interjecting additional information where necessary. If the scoring director identified a particular problem that seemed to be causing difficulty across teams, he or she discussed it roomwide to help ensure that everyone was hearing the same explanation. TEA staff monitored this entire process.

Like team leaders, readers had to demonstrate accuracy in their scoring before they could begin reading packets of responses. Readers were allowed three opportunities to qualify. Any reader unable to meet the standards set by TEA was dismissed.

Training of the analytic readers for all grades and of 1/2 score verification specialists for the exit level followed a similar pattern, except that the training was performed by the respective coordinators.

ONGOING ROOMWIDE TRAINING

After the initial training, ongoing training was provided routinely to prevent “drift” and to ensure high reader agreement. Scoring directors planned for at least three ongoing training sessions a week. These methods are described in the following paragraphs.

One method was the scoring and discussion of sets of up to five papers each. These papers were selected by team leaders while spot-checking readers. The sets were a mix of all the score points and included some borderline responses. The scores of these papers were agreed on both by the scoring directors and by the team leaders. Both shifts used these sets. Discussion of these sets sometimes occurred roomwide and sometimes in teams.

While scoring papers and spot-checking the accuracy of readers’ scoring, team leaders were instructed to collect various types of problematic papers. These papers were reproduced and put into small sets for readers to score. After both scoring directors, the project monitor, and, in the case of a “decision” paper, a TEA representative agreed on the scores of these papers, the sets were administered to the readers. Discussion of these papers was conducted roomwide. Only one or two of these sets were needed, depending on the grade level. If individuals needed more help, the floating team leader worked with them.

Once the scoring of papers was underway, the scoring directors started accumulating copies of papers that were typical close, or “line,” calls. The scoring directors reviewed these papers with senior scoring staff and then circulated them among team leaders to ensure team-to-team consistency on these difficult decisions. Every Monday the scoring directors reviewed the rubrics with readers and had them reread their anchor papers, emphasizing any area that appeared to be giving readers problems.

MONITORING OF INDIVIDUAL READERS

In addition to the ongoing training methods mentioned above, the scoring centers employed a number of informal methods to identify individual reader scoring problems. Scoring directors and team leaders relied on individual and small group retraining to ensure that readers were consistently applying the preestablished criteria when scoring. Team leaders spot-checked and annotated reader packets throughout the project and then returned packets to the readers for their review. If necessary, the team leaders would provide one-on-one assistance to a reader and discuss discrepant scores. Readers also flagged papers that were difficult for them to score. Team leaders read these papers and then discussed each paper with the reader who had flagged it.

Early in the project, team leaders closely monitored all readers, spot-checking according to the following: scoring trends identified from training results, reports of “true” score reliability (see “Validity Packets” on p. 75), and daily reader status reports (see “Data-Entry Procedures and Resulting Reports” on p. 77). The need to spot-check every reader decreased as it became clear which readers were consistently applying the scoring criteria and which needed additional support. At this point, team leaders concentrated on readers who scored below 80% on the validity packets and/or who were below the room average on the daily reader status reports. They conducted hands-on retraining by identifying problem papers, having readers articulate their reasoning for assigning a particular score, and reinforcing the rubric and training papers to improve readers’ accuracy.

Another method used when a team leader suspected that a certain reader might not be using the criteria properly was to obtain a regular packet that had been scored first by the floating team leader. Distribution of this type of packet was done routinely so attention would not be called to it as a training device. The reader’s team leader then compared the floating team leader’s scores with those of the reader. If there were a number

of discrepant scores, the floating team leader or the reader's team leader discussed the papers with the reader to help him or her apply the criteria consistently.

Packets scored by a reader identified as having difficulty applying the criteria were retrieved and rescored by his or her team leader or by a reader at or above room average. The team leader then discussed with the reader the papers that had received discrepant scores. Any reader who could not be successfully retrained on the criteria was dismissed.

VALIDITY PACKETS

As a method of detecting whether a room of readers was drifting from the scoring criteria, packets of prescored student responses, called validity packets, were assembled for each grade. For each scoring session, these papers were chosen from packets that had been read and agreed on by two team leaders, the appropriate scoring director(s), and the TEA representative(s). "True" scores were assigned to these papers.

For the spring 2002 scoring project, nine validity packets were used at exit level, each containing fifteen randomly mixed papers representing all score points. Six validity packets containing fifteen papers each were used for Grade 4 English and Grade 8; three validity packets were used for Grade 4 Spanish. During the day, for the first two weeks of the project, each packet received first and second readings both in the morning and in the afternoon. After the second week of scoring, each packet received first and second readings once each day. Because the evening shift had fewer hours and fewer readers, the packets were first- and second-read once each evening. All readers read one or more of these validity packets during the course of the scoring project.

For the October 2001 and July 2002 exit level retests, time constraints required some procedural modifications. Five validity packets, which consisted of ten papers each, were first- and second-read in the morning and again in the afternoon so that all readers read at least one packet during the scoring session. Since the May 2002 scoring session lasted only one week, no validity packets were used during that session.

For each validity packet NCS Pearson printed multiple monitor sheets listing each composition's unique preprinted identification number. At the end of each shift, completed monitor sheets were processed, reports were printed, and new monitor sheets were inserted for the next shift's scoring. Thus, senior staff had access to validity packet reports twice daily and could detect room drift and/or reader drift almost as soon as it began.

ANALYTIC SCORING

At all grades each composition that receives either a rating of 0 or a score of 1 is evaluated analytically to provide information about the specific weaknesses that cause it to be unsuccessful. Analytic readers are trained on all the analytic features simultaneously. Papers that exemplify the range of unsuccessful compositions and that TEA and senior scoring center staff agree on in advance are selected as training papers. The scoring director first reads and discusses the guide with the analytic readers. The guide includes ten sample papers that are chosen to represent a variety of analytic scores. The analytic readers read, score, and discuss four additional sets of ten papers each. The readers begin "live" scoring when they are able to demonstrate accuracy on all analytic categories.

EXIT LEVEL SCORE VERIFICATION

Since the spring 1992 exit level scoring session, TEA's contractors have used a score-verification procedure to further evaluate all responses that received a 1 during the holistic scoring process. A special team of readers are trained exclusively on the 1/2 line by using the 1s and 2s from the holistic guide and split set, along with

ten-paper sets that make up the rest of the 1s and 2s from the holistic sets and qualifying rounds. Additional responses found in the live papers (approved by the scoring director, the coordinator of the analytic readers and 1/2 score verification specialists, and TEA staff) are selected to round out the training sets. If any response scored by a member of the specialist team is thought to be higher than a 1, it is read by the specialist coordinator. If the coordinator agrees, the response is then read by the scoring director. If the scoring director also agrees, the score is changed; if he or she disagrees, the response is read by the project monitor, who makes the final decision unless it involves an issue that should be brought to the attention of TEA. In that case the response is sent to TEA for a final scoring decision.

NONSCORABLE RESPONSES

During holistic scoring, if a reader believes that a paper may be nonscorable, the paper is flagged for the scoring director to read and score. If the scoring director finds it to be nonscorable, the second reading is performed independently by the other scoring director or by the project monitor. Nonscorable responses are then evaluated by the analytic readers.

PROCEDURES

PAPER-FLOW AND RESOLUTION PROCEDURES

A scoring director supervised the day shift of readers for each grade; his or her counterpart supervised the evening shift. Two scoring directors were used for each shift during the spring 2002 exit level project. Continuity between the day shift and the evening shift was maintained in a number of ways, including a 2 1/2-hour overlap in the work schedule of the scoring directors. The schedules of supervisors in the data-entry room and warehouse overlapped so that continuity could be maintained in those areas.

The logistics of paper flow in the scoring centers were carefully planned and carried out. The answer documents were sent to the NCS Pearson TechRidge facility in Austin, where they were scanned. During the scanning process, the two lined pages on which students wrote their compositions were separated from the multiple-choice section of the answer document. The two sections of the answer document were linked by a unique number printed on each page so that the composition's score could be added to the student's record once scoring was complete. The writing pages were then assembled into packets containing 40 or fewer papers each. A packet header sheet was placed with the packet of papers, and the packet was stapled together and put into an envelope with two scoring monitor sheets. As a result of this process, the only identifying information on the student papers was the six-digit identification number preprinted on the answer document. Unless students signed their names, wrote about their hometowns, or in some way provided other identifying information, the readers had no knowledge of whom the students were. The unavailability of identifying information on the papers helped ensure unbiased scoring.

The packets were then boxed by grade and shipped to the scoring centers in Austin, Dallas, and Lawrence. Whenever a scoring room needed additional papers, an aide carried packets to the room. The room aide and team leaders handled all paper flow in the scoring rooms.

Each reader independently read an entire packet of papers, writing his or her reader number on both the packet envelope and the monitor sheet. The reader recorded the scores for the papers on the monitor sheet, on which the identification numbers of the essays in the packet had been preprinted. The completed first-reading monitor sheet was separated from the packet envelope before the packet was given to a second reader. The reader number on the packet envelope identified the reader's team as well as the individual to ensure that the same packet would not be read by another reader on the same team as the first reader.

Following scanning of both the first- and second-reading monitors, third-reading monitor sheets identifying responses needing an additional reading were produced. Only readers identified as being above room average in the accuracy of their scoring were allowed to be resolution, or third, readers. Early in the project they were selected on the basis of their performance in training, such as their scores on training sets and the caliber of their questions and comments, along with their team leaders' assessment of their "live" scoring. Later the daily reader status reports and validity reports were invaluable in identifying the readers whose scoring accuracy was above room average. Designated third readers were not allowed to score third readings exclusively. Rather, they were required to score at least two 40-paper packets daily so that sufficient data could be collected to monitor their scoring on an ongoing basis. Any third reader whose perfect agreement rate on the daily status report dropped was confined to performing first and second readings. Occasionally a fourth reading of a student paper was necessary. When this occurred, the fourth-reading monitor sheets were matched to the packets and given to scoring directors for scoring.

Responses requiring analytic scores and exit level verification of 1/2 line scores were identified on an analytic monitor sheet. A 1/2-line monitor sheet was matched with the appropriate packet and delivered to the analytic and 1/2-line scoring room.

DATA-ENTRY PROCEDURES AND RESULTING REPORTS

The packet monitor sheets were scanned at the scoring centers, and the scores were transmitted to NCS Pearson in Iowa City. After the scores for the first and second readings of a packet had been scanned, the resolution monitor sheet (third-reading monitor) was produced. NCS Pearson transmitted the data for third-reading monitor sheets (as well as fourth-reading, analytic, and specialist monitor sheets) to the NCS Pearson Performance Scoring Center's printer. The monitors were then printed and delivered to the warehouse.

The data also informed project status reports that gave senior staff and scoring directors up-to-date information on the progress of the entire project at all scoring centers. These reports provided a wealth of information about the scoring patterns of individual readers. In addition to the number of responses read by each reader, the reports included the following for each reader: number of third readings completed, percentage of responses read in perfect agreement with the other scorer, and percentage of responses read in perfect agreement with the other scorer in combination with responses read in perfect agreement with the resolver. In every resolution reading, one reader's score was judged to be incorrect; consequently, the reports had three adjacent score categories, 1/2, 2/3, and 3/4. These showed the number of times the reader's incorrect scores were higher and/or lower for each of the adjacent score categories. The final columns on the reader status reports gave the readers' distribution of score points— that is, what percentage of a particular reader's scores were 1s, 2s, etc. Accompanying the daily (or current) reader status report was the year-to-date report, which had the same information but was cumulative for the project as of that date.

SCORE APPEALS

NCS Pearson rescores any TAAS written composition about which questions have been raised regarding the assigned score. Through a telephone call to the district contact person, NCS Pearson provides an individual analysis of the composition in question.